Journal of Multivariate Analysis 100 (2009) 2224–2236

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Nearest neighbor conditional estimation for Harris recurrent Markov chains

Alessio Sancetta *

Deutsche Bank, London, UK

ARTICLE INFO

Article history:
Received 1 April 2008
Available online 8 July 2009

AMS subject classifications:
62G05
62G08
62M05

Keywords:
Markov chain
Nonparametric estimation
Semiparametric estimation
Sequential forecasting

ABSTRACT

This paper is concerned with consistent nearest neighbor time series estimation for data generated by a Harris recurrent Markov chain on a general state space. It is shown that nearest neighbor estimation is consistent in this general time series context, using simple and weak conditions. The results proved here, establish consistency, in a unified manner, for a large variety of problems, e.g. autoregression function estimation, and, more generally, extremum estimators as well as sequential forecasting. Finally, under additional conditions, it is also shown that the estimators are asymptotically normal.

1. Introduction

This paper is concerned with conditional nonparametric and semiparametric estimation from data generated by a stochastic process that can be represented as a Harris Recurrent Markov Chain (HRMC). The class of HRMC is fairly general and includes processes that may not be stationary (e.g. univariate random walks). The basic interest of the paper is to consider a process $X = (X_i)_{i\in\mathbb{N}}$ with values in some set E (with a partial order $\leq$) and some measurable function f on E and to estimate $\mathbb{E}_{i-1}f(X_i)$ ($\mathbb{E}_{i-1}$ is expectation conditional on the sigma algebra generated by $(X_s)_{s<i}$) or some related quantity like $\inf_f \mathbb{E}_{i-1}f(X_i)$ over some class of functions from which we can derive conditional extremum estimators. Most common examples include conditional distribution function estimation ($f(x) = I\{x \leq y\}, y \in E$), regression estimation ($f(x) = x$) and, as just mentioned, conditional extremum estimators. The goal is to validate, in a unified manner, the application of nearest neighbor estimation to a general class of time series problems. An example of such problems is conditional likelihood estimation.

Assuming the HRMC condition, the goal is to state simple general conditions that would imply consistency, and in some cases asymptotic normality, for nonparametric and/or semiparametric estimation, avoiding mixing conditions. Mixing conditions are commonly used in the nonparametric literature (e.g. [1], for an early reference, [2], for regression function estimation of functional data), though recently, more general weak dependence conditions have also been employed [3]. When the hypothesized data generating process (DGP) is available, computation of mixing conditions is difficult [4] and for this reason weak dependence coefficients are used ([5,6] for a review).

As an alternative to mixing and dependence coefficients, we may suppose that the data come from a given class of stochastic processes, but no other information is available. We may not even know if the process is stationary. The natural

* Corresponding address: Via Flaminia Nuova 213, 00191 Rome, Italy.
E-mail address: asancetta@gmail.com.
URL: http://www.sancetta.googlepages.com.

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2009.06.013

question to ask is the following: is it possible to identify a broad class of stochastic processes in which many discrete time statistical models can be embedded and such that nonparametric estimation is still consistent? This question has been positively answered by Yakowitz [7], where a slightly more general class of stochastic processes than HRMC has been considered, but attention is limited to autoregression function estimation. Karlsen and Tjostheim [8] slightly restricted the class of stochastic processes, but considered more general nonparametric estimation problems. Karlsen and Tjostheim [8] studied nonparametric kernel estimation, while [7] used a nearest radius approach, also used here and to be described in due course. The nearest radius approach considerably simplifies the argument. Markov chains (MC) and in particular HRMC have also been considered as an important case of DGP around which to develop empirical methods for inference (e.g. [9,10].

The present paper focuses mainly on weak sufficient conditions that assure consistency for a variety of estimators. However, under more restrictive conditions central limit theorems can also be inferred and details are provided. Inferential arguments in conditional nonparametric estimation have also been carefully handled by Karlsen and Tjostheim [8]. Restricting our interest to consistency only, the conditions used here are particularly simple. Unlike [7], this paper is not restricted to autoregression function estimation, but more general nonparametric and semiparametric procedures are studied. The main idea is to be able to consistently estimate the conditional distribution function. This allows us to derive consistency for a large number of nonparametric and semiparametric problems, imposing mild smoothness conditions on the transition distribution only. Applications will be discussed and include local conditional likelihood estimation. In this respect, the class of problems considered includes extremum estimators, hence, it is more general than some of the problems considered by Karlsen and Tjostheim [8]. Moreover, these authors consider nonparametric estimation for real valued HRMC, though based on their theoretical results, this condition could be relaxed. Here we shall consider a more general state space E. To the author's knowledge, this is the first study that considers consistency for conditional extremum estimators in this general framework.

Section 2 discusses the nearest neighbor procedure and states minimal conditions under which the nonparametric estimator of the conditional distribution function is consistent. This result is then used to show consistency for a variety of problems. Conditions that imply asymptotic normality of the estimators are derived. Further discussion about the results can be found in Section 3. Proofs of results can be found in Section 4. Next we just mention a few models that can be embedded in HRMC.

1.1. Many important econometric and statistical models are HRMC

Recall that an MC is a discrete time process such that, conditioning on the present, the future and the past are independent. Then, an HRMC, say X, with state space E is an irreducible MC such that

$$\Pr\left(X_n \in B i.o. | X_0 = x\right) = 1, \quad x \in E \text{ (}i.o.\text{ stands for infinitely often)}$$

for any set B of positive ψ measure, where ψ is some suitable sigma finite measure (e.g. [11], for details).

By a suitable definition of the state space E, it is possible to embed many stochastic processes in the class of HRMC, under suitable restrictions (e.g. non-explosive coefficients). Linear autoregressive, SETAR, multilinear, and ARCH models, all fall within the class of HRMC. Many examples can be obtained by considering the class of models that can be embedded in the following multivariate stochastic difference equation

$$X_n = A_n X_{n-1} + B_n, \tag{1}$$

where $(A_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ are iid matrix and vector random variables ([12], for details on recurrence and references). ARCH models of finite order are an example of models that can be embedded in (1) (e.g. [13]). Further details on examples can be found in [11], ch. 2).

2. Conditional estimation using nearest neighbors

Let $X = (X_n)_{n\in\mathbb{N}}$ be an aperiodic HRMC on a countably generated state space $(E, \mathcal{E})$ with transition probability $P(x, A)$ and invariant measure π. The measure induced by the transition kernel P at $x \in E$ is denoted by π_x, i.e. $\pi_x(A) := P(x, A)$, $A \in \mathcal{E}$. The Markovian probability with initial value x is denoted by P_x, i.e. $P_x(X_n \in A) = \Pr(X_n \in A | X_0 = x)$, $A \in \mathcal{E}$. We shall use linear functional notation, as is commonly done in the MC literature, e.g. for some suitable function f, $Pf(x) := \int_E f(y) P(x, dy)$ and for some set $B \subset E$, $Pf(B) := \int_B \int_E f(y) P(x, dy) [\pi(dx)/\pi(B)]$ (and the use of this notation will not require further explanation). Note that if $\pi(E) < \infty$ the HRMC is said to be positive recurrent, while null recurrent if $\pi(E) = \infty$. Null recurrent MC do not possess stationary distribution. At first, we shall be concerned with estimation of

$$P(x, \{y \in E : y \le s\}) = \Pr(X_n \le s | X_{n-1} = x),$$

where we assume that E is a partially ordered set, e.g. inequalities are meant elementwise and the meaning of this notation will be assumed throughout without reminder. By relatively standard results, consistent estimation of the transition distribution allow us to derive in a unified manner a wide variety of estimators which are discussed in the sequel.

For simplicity, but with abuse of notation, we shall write $P(s|x)$ as a short cut for $P(x, \{y \in E : y \le s\})$, the conditional distribution function. Finally, $\asymp$ and $\lesssim$ will be used to denote equality and inequality up to a finite absolute multiplicative constant.

2.1. The estimator

We shall generalize [7] allowing E to be a state space more general than $\mathbb{R}$. Denote by $m \to \infty$ the number of neighbors. The estimator is derived in terms of the recurrence times of X to some conditioning set $B(x, r_m) \to \{x\}$ as $r_m \to 0$, which is a ball of d-radius r_m. Hence, we suppose that E is metrizable by some metric d. When $E \subseteq \mathbb{R}^K$ ($K \geq 1$ but finite in the sequel), d is topologically equivalent to the Euclidean distance. To ease notation, we shall use B_m, $B_m(x)$ and $B(x, r_m)$ interchangeably, whichever is felt to be more appropriate. For any set $B \subseteq E$, define $T_B := \inf\{n > 0 : X_n \in B\}$ and $T_B(i) := \inf\{n > T_B(i-1) : X_n \in B\}$, $T_B(1) := T_B$, i.e. $T_B(i)$ is the time of the ith visit to B. Hence,

$$\hat{P}_m(s|B_m) := \hat{P}_m(B_m, \{y \in E : y \leq s\}) = \frac{1}{m}\sum_{i=1}^{m} I\left\{X\left(T_{B_m}(i) + 1\right) \leq s\right\} \tag{2}$$

is an m nearest neighbor estimator for the one step ahead conditional distribution ($X(i) = X_i$ for typographical reasons) based on a sample of (random) size $\left(T_{B_m}(m) + 1\right)$. The same linear functional notation used for P will also be used for $\hat{P}_m$, e.g. $\hat{P}_m f(B_m) = \int_E f(y)\hat{P}_m(B_m(x), dy)$.

Note that by the Harris recurrence assumption, $T_B(i) < +\infty$ a.s. for each i. This means that, as $n \to \infty$, we shall be able to allow $m \to \infty$ so that the estimation error goes to zero. However, for consistency, we shall also require $B(x, r_m) \to \{x\}$ so that the bias is vanishing (i.e. the conditioning set needs to shrink as the sample size increases). To this end, we shall first fix a sequence $r_m \to 0$ as $m \to \infty$. This means that having fixed a radius r_m, we shall wait for m visits to $B_m(x)$ in order to construct $\hat{P}_m$, which is an m nearest neighbor estimator. By the Harris recurrence, this will happen a.s. in finite time for any m.

Let $L(n)$ be a slowly varying function of n at infinity (e.g. [14]). If we assume X to be β-recurrent (using the terminology in [8]), then, by Theorem 2.1 in [15], $\sum_{i=1}^{n} f(X_i) \asymp n^{\beta}L(n)$ in probability, $\beta \in [0, 1]$, for any non-negative π integrable f such that $\pi f > 0$. (Note that [15], calls this MC regular and expresses the condition in terms of recurrent times of D-sets: using results about atoms and small functions, the two definitions are equivalent, e.g. [15].) Clearly, $\beta = 1$ is the positive recurrent case. It is well known (e.g. [15,8]) that a random walk is recurrent of index $\beta = 1/2$. Hence, if we knew β, we would know that $n^{\beta}/m_n \to \infty$ is necessary. (When $\beta = 1$, we recover the familiar necessary condition for consistency on the m neighbors.) Mutatis mutandis, this is the approach of [8], though the formal approach requires the use of Nummelin splitting technique (e.g. [11]) and considerable technicalities. Note that in a number of results proved by Karlsen and Tjostheim [8] the bandwidth depends of β or some other unknown quantity (e.g. their assumption A5 and results in their Section 5). Here, no assumption of regularity is made so that the estimator can be constructed only using the predetermined sequence of sets $B(x, r_m)$. Noting that $\pi(B(x, r_m)) < \infty$ because π is sigma finite, under the assumption of β recurrence in Karlsen and Tjostheim [8], we could use Theorem 2.1 in [15] and impose conditions directly on the neighbors, without worrying about the choice of the radius r_m. Clearly, this would require a knowledge of β. For example, by Theorem 2.1 and Eq. (1.2) in [15] infer that

$$m = \sum_{i=1}^{n} I\{X_i \in B\} \asymp n^{\beta}L(n)\pi(B) + o\left(n^{\beta}L(n)\right)$$

in probability. Suppose that $E = \mathbb{R}^K$. Then from the proof of Lemma 2.2 in [30] deduce that $n^{\beta}L(n)\pi(B) \to \infty$ if $n^{\beta}L(n)r^K \to \infty$. Hence, we are able to relate $m = m_n$ to $r = r_m$. For the case $\beta = 1$ and $L(n) = 1$, we recover the usual condition for convergence of the nearest neighbor estimator for iid random variables: $r \to 0$ and $nr^K \to \infty$ (e.g. [30]). Note that for consistency all we need is $m \to \infty$, $B \to \{x\}$ (just use the same martingale argument as in [7]).

2.2. Consistency of the conditional empirical distribution function

The conditions used for consistency of the conditional empirical distribution are formally listed below. Further conditions might be required in the applications and these will be stated when needed.

Condition 1. *$X := (X_n)_{n\in\mathbb{N}}$ is an aperiodic Harris recurrent Markov chain on a state space $(E, \mathcal{E})$ with countably generated sigma algebra $\mathcal{E}$, and with transition probability $P(x, A)$ and invariant measure π. E has a partial order $\leq$ and is equipped with a metric d.*

Condition 2. $\Pr(X_1 \leq s|X_0 = x)$ *is a.s. continuous in $x \in E$ for any $s \in E$.*

Remark 3. By the Lebesgue Differentiation Theorem, if continuity does not hold, the results are still true for π-almost all x when $E \subseteq \mathbb{R}^K$ (see the proof of Lemma 32 for details).

Condition 4. *$m \to \infty$ and $r_m \to 0$.*

Remark 5. By Conditions 1 and 4 is always feasible.

Theorem 6. *Under Conditions 1, 2 and 4,*

$$\hat{P}_m\left(s|B_m\left(x\right)\right) \xrightarrow{a.s.} P\left(s|x\right)$$

pointwise, and if $E \subseteq \mathbb{R}^K$,

$$\sup_{s\in E}\left|\hat{P}_m\left(s|B_m\left(x\right)\right) - P\left(s|x\right)\right| \xrightarrow{a.s.} 0.$$

Theorem 6 shows that for a general state space (countably generated and metrizable), the convergence holds pointwise a.s. If we restrict attention to $E \subseteq \mathbb{R}^K$, the convergence holds uniformly a.s. We now use this result to consider applications to statistical estimation problems.

2.3. Estimation of conditional minimum estimators

The following set up is abstract but an application is discussed in Section 3.

Consider the following problem

$$\inf_{f\in\mathfrak{F}} Pf\left(x\right)$$

where $\mathfrak{F}$ is some set of functions with values in $\mathbb{R}$ (and recall that $Pf\left(x\right)$ is the expectation of $f\left(X_n\right)$ conditioning on $X_{n-1} = x$). Suppose $f\left(y\right) = f_\theta\left(y\right)$ is convex in $\theta \in \Theta$ for some suitable set Θ. Then, the above problem can be seen as an abstract version of the more common problem of minimizing the risk $Pf_\theta\left(x\right)$ with respect to θ. Solution of this problem allows us to define population values for many statistical estimators.

Example 7. Suppose $f_\theta\left(y\right) = |y-\theta|^2$ and $y \in E \subseteq \mathbb{R}$. Then,

$$\arg\inf_{\theta\in\Theta} Pf_\theta\left(x\right) = \mathbb{E}\left(X_n|X_{n-1} = x\right),$$

i.e. the expectation of X_n conditioning on $X_{n-1} = x$.

Example 8. Suppose $f_\theta\left(y\right) = u\left|y-\theta\right|^+ + \left(1-u\right)\left|y-\theta\right|^-$ and $y \in E \subseteq \mathbb{R}$, $u \in (0, 1)$. Then,

$$\arg\inf_{\theta\in\Theta} Pf_\theta\left(x\right) = Q_u\left(X_n|x\right),$$

which denotes the u quantile of X_n conditioning on $X_{n-1} = x$.

For a general treatment of the problem, it is simpler to define minimization with respect to $f \in \mathfrak{F}$ rather than $\theta \in \Theta$. We need to restrict the class of functions $\mathfrak{F}$ to be considered.

Condition 9. *For any* $x \in E$*, the following holds:*

i. $\mathfrak{F}$ *has a measurable envelope function* $F := \sup_{f\in\mathfrak{F}}|f|$ *such that* $\limsup_m PF^p\left(B_m\left(x\right)\right) < \infty$ *for some* $p > 1$;
ii. $\mathfrak{F}$ *is a family of* π_x*-a.s. equicontinuous functions on* E.

Remark 10. A family of equicontinuous functions contains functions that are not necessarily Lipschitz for a given metric, e.g. any finite set of continuous functions.

Remark 11. If $E \subseteq \mathbb{R}^K$, we can allow for more general families of functions, possibly discontinuous. To limit the notational burden in the text, we do not discuss this special case, but details can be found in Section 3.

Corollary 12. *Under Conditions 1, 2, 4 and 9,*

$$\sup_{f\in\mathfrak{F}}\left|\hat{P}f\left(B_m\left(x\right)\right) - Pf\left(x\right)\right| \xrightarrow{a.s.} 0,$$

for any $x \in E$.

Remark 13. This result is a generalization of Theorem 2 in [7], where, mutatis mutandis, $p > 2$ is required. Moment conditions higher than 2 are also used for consistency in Theorem 5.2 of [8], though their results are not directly comparable because they use a different nonparametric estimator. Note that these authors do not consider the uniform in $\mathfrak{F}$ case.

The above result can be used to derive conditional extremum estimators. Define

$$\hat{f}_m(x) := \arg\inf_{f\in\mathfrak{F}} \hat{P}_m f(B_m(x)) \quad \text{and} \quad f_0(x) := \arg\inf_{f\in\mathfrak{F}} Pf(x),$$

so that f_0 is the unfeasible optimal choice of $f \in \mathfrak{F}$ (i.e. unknown), while $\hat{f}$ is the feasible estimator. Then, under an additional identifiability condition, we have that $\hat{f}$ and f are close to each other for each fixed x. To formalize this we need the following additional condition, which is minimal.

Condition 14. *For any $x \in E$, let $G = G_x$ be any arbitrary open set that contains $f_0(x)$ and let G^c be its complement. Then,*

$$\inf_{f\in G^c} Pf(x) > Pf_0(x).$$

Corollary 15. *Suppose $(\mathfrak{F}, \rho)$ is a metric space. Under Conditions 1, 2, 4, 9 and 14,*

$$\rho\left(\hat{f}_m(x), f_0(x)\right) \xrightarrow{p} 0,$$

for any $x \in E$.

2.4. Sequential forecasting

We now consider sequential forecasting. Define

$$\hat{f}_{m,n} := \hat{f}_m(X_{n-1}) \quad \text{and} \quad f_n := f(X_{n-1}),$$

so that f_n is the unfeasible $\mathcal{F}_{n-1}$ measurable optimal choice of $f \in \mathfrak{F}$, while $\hat{f}_{m,n}$ is the feasible $\mathcal{F}_{n-1}$ measurable estimator ($\mathcal{F}_{n-1}$ is the sigma algebra generated by $(X_s)_{s<n}$). The measurability is required for the forecasts to be admissible (i.e. no use of future information is allowed). The goal is to apply Corollary 15 to the problem of sequential forecasting.

Theorem 16. *Suppose $(\rho, \mathfrak{F})$ is a metric space and $\rho\left(\hat{f}_{m,n}, f_n\right)$ is asymptotically P_x-uniformly integrable in m for any n. Under Conditions 1, 2, 4, 9 and 14,*

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_x \rho\left(\hat{f}_{m,n}, f_n\right) \to 0,$$

where $\mathbb{E}_x(X_n) = \mathbb{E}(X_n|X_0 = x)$, i.e. expectation w.r.t. P_x.

Theorem 16 says that the average loss incurred using the estimated forecast $\hat{f}_{m,n}$ is equivalent to the one incurred using the optimal unfeasible sequential forecast f_n.

To provide some understanding of the condition "$\rho\left(\hat{f}_{m,n}, f_n\right)$ is P_x-uniformly integrable for any n" suppose: $f_n := \mathbb{E}_{n-1}X_n$, X is a random walk with values in $\mathbb{R}$ and $\rho(x, y) = |x - y|$. Then, $\hat{f}_{m,n} := \sum_{i=1}^{m} X(T_B(i) + 1)/m$ where $B = B_m(X_{n-1})$ and $\mathbb{E}_x\left|\hat{f}_{m,n} - f_n\right|^p < \infty$ under a $p > 1$ moment condition on the innovations of the random walk. Hence, $\rho\left(\hat{f}_{m,n}, f_n\right)$ is P_x-uniformly integrable in m for any n. We now turn to conditions that allow us to derive the asymptotic distribution of the nearest neighbor estimator.

2.5. Asymptotic normality

Strengthening Conditions 2 and 4 we can establish asymptotic normality of the nearest neighbor estimators of $Pf(x)$.

Condition 17. *For any positive real r and $x, x' \in E$, such that $d\left(x, x'\right) \le r$,*

$$\left|Pf(x) - Pf\left(x'\right)\right| \lesssim r^{\alpha}$$

with $\alpha > 0$.

Condition 18. *$m \to \infty$ and $r_m \to 0$ such that $\sqrt{m}r_m^{\alpha} \to 0$.*

The above two conditions allow us to control the bias in the procedure. The next is slightly stronger than needed, but simple.

Condition 19. *Let $\mathbb{E}_{T(i)}$ be expectation conditionally on $X\left(T_{B_m}(i)\right), X\left(T_{B_m}(i)-1\right), \ldots, X(0)$. Let $\mathfrak{F}$ be a finite set of of uniformly bounded functions from E to $\mathbb{R}$. Define*

$$\sigma_{m,x}(f,g) := \frac{1}{m}\sum_{i=1}^{m}\left[\left(1-\mathbb{E}_{T(i)}\right)f\left(X\left(T_{B_m}(i)+1\right)\right)\right]\left[\left(1-\mathbb{E}_{T(i)}\right)g\left(X\left(T_{B_m}(i)+1\right)\right)\right].$$

Then,

$$\lim_{m\to\infty}\sigma_{m,x}(f,g) = P(fg)(x) - Pf(x)\,Pg(x), \quad f,g\in\mathfrak{F},$$

in probability.

A central limit theorem (CLT) can be obtained.

Theorem 20. *Let $\mathfrak{F}$ be a finite set of of uniformly bounded functions from E to $\mathbb{R}$. Under Conditions 1 and 17–19, for any $x\in E$,*

$$\left(\sqrt{m}\left[\hat{P}f\left(B_m(x)\right) - Pf(x)\right]\right)_{f\in\mathfrak{F}} \to \left(\mathbb{G}_x(f)\right)_{f\in\mathfrak{F}}$$

in distribution, where $\left(\mathbb{G}_x(f)\right)_{f\in\mathfrak{F}}$ is a centered Gaussian vector with covariance matrix

$$\mathbb{E}\mathbb{G}_x(f)\,\mathbb{G}_x(g) = P(fg)(x) - Pf(x)\,Pg(x), \quad f,g\in\mathfrak{F}.$$

Letting $f(y) = I\{y\le s\}$ the result applies to the conditional empirical distribution function where the covariance matrix is given by

$$\mathrm{Cov}\left(I\{X_1\le s\}, I\left\{X_1\le s'\right\}|X_0=x\right) = P(s|x)\wedge P\left(s'|x\right) - P(s|x)\,P\left(s'|x\right).$$

The above result only holds when $\mathfrak{F}$ is a finite set. Further discussion of the above results, together with extensions is provided in Section 3.

3. Discussion

Next we give a simple application of the previous results. We then provide some short discussion about some examples of general state space. We show that we can considerably improve on Condition 9 ii. when $E\subseteq\mathbb{R}^K$. Finally, we provide details on how to deduce a uniform CLT.

3.1. Conditional likelihood estimation

Suppose that the transition kernel admits the following representation

$$P(x,A) = \int_A p(y;\theta(x))\,\mu(\mathrm{d}y),$$

where μ is a sigma finite measure and $\theta(x)$ is a function of x taking values in Θ. Then, $(p(y;\theta))_{\theta\in\Theta}$ is a model where $\theta = \theta(x)$ is unknown and we ignore a parametric form for $\theta(x)$. Hence the model $p(y;\theta(x))$ depends on the infinite dimensional parameter $\theta(x)$.

Example 21. Suppose $X_n = \theta(X_{n-1})Z_n$, where $(Z_n)_{n\in\mathbb{N}}$ is iid standard Gaussian noise and $\theta(X_{n-1})$ is a function of X_{n-1}. Then, $p(y;\theta(x)) = \phi(y/\theta(x))/\theta(x)$ denoting the standard Gaussian density by ϕ. This is a simple Markovian model for heteroskedastic data. If we are unable or unwilling to make a parametric assumption for $\theta(x)$, then, we could use nonparametric methods to estimate it. The conditionally Gaussian ARCH process of finite order is a special fully parameterized case of this model.

In some models (notably the ones belonging to the exponential family), we also have that there is a function g such that

$$\theta(x) = \int_A g(y)\,p(y;\theta(x))\,\mu(\mathrm{d}y).$$

Example 22. Suppose $p(y;\theta) = \exp\{\langle a(\theta), g(y)\rangle + b(\theta)\}\,c(y)$, for some positive functions a, b and c, where $\theta = \int g(y)\,p(y;\theta)\,\mathrm{d}\mu(y)$, and $\langle\bullet,\bullet\rangle$ is the inner product. Clearly, a and g could be vector valued functions. This density is said to belong to the exponential family model, with natural parameter θ, canonical parameter $a(\theta)$ and canonical statistic $g(x)$ (e.g. [16]). The Gaussian, the Poisson and the Binomial distributions all belong to this family.

When $p\left(y;\theta\left(x\right)\right)$ is the density kernel, it is natural to ask if nonparametric estimation can be used to consistently estimate $p\left(y;\theta\left(x\right)\right)$ or $\theta\left(x\right)$. Clearly, the case

$$\theta\left(x\right)=Pg\left(x\right)=\int_{A}g\left(y\right)p\left(y;\theta\left(x\right)\right)\mu\left(\mathrm{d}y\right)$$

is dealt by Corollary 12. A general alternative to this method is to choose $\theta\left(x\right)$ to maximize

$$\mathbb{E}\left[\ln p\left(X_n;\theta\right)|X_{n-1}=x\right] \tag{3}$$

with respect to θ. Denoting the true unknown function to estimate by $\theta_0\left(x\right)$, the justification of (3) is the usual one via the scoring rule: under regularity conditions,

$$\begin{aligned}\left(\partial/\partial\theta\right)\mathbb{E}\left[\ln p\left(X_n;\theta\right)|X_{n-1}=x\right]&=\int_{E}\frac{\left(\partial p\left(y;\theta\right)/\partial\theta\right)}{p\left(y;\theta\right)}p\left(y;\theta_0\left(x\right)\right)\mu\left(\mathrm{d}y\right)\\&=\int_{E}\left(\frac{\partial p\left(y;\theta_0\left(x\right)\right)}{\partial\theta_0\left(x\right)}\right)\mu\left(\mathrm{d}y\right)=0\end{aligned}$$

if $\theta=\theta_0\left(x\right)$. Corollary 12 shows that, under regularity conditions,

$$\sup_{\theta\in\Theta}\left|\int_{E}\ln p\left(y;\theta\right)P_m\left(\mathrm{d}y|B_m\left(x\right)\right)-\mathbb{E}\left[\ln p\left(X_n;\theta\right)|X_{n-1}=x\right]\right|\xrightarrow{a.s.}0, \tag{4}$$

so that the semiparametric likelihood approach is consistent: this is just an application of Corollary 15. In particular the following is easily verified.

Corollary 23. *Suppose* $\ln p\left(y;\theta\right)$ *is uniformly continuous in* $\theta\in\Theta$*, and, for some* $p>1$, $\sup_\theta\left|\ln p\left(y;\theta\right)\right|$ *is in* $L_p\left(\pi_{x'}\right)$ *for any* x' *in a neighborhood of* x*. Suppose* $\left(\Theta,\rho\right)$ *is a totally bounded metric space. Then, under the conditions of Theorem 6, (4) holds. Moreover, if* $\ln p\left(y;\theta\right)$ *has a unique maximum, then*

$$\rho\left(\hat{\theta}_m\left(x\right),\theta_0\left(x\right)\right)\xrightarrow{p}0$$

where

$$\hat{\theta}_m\left(x\right):=\arg\inf_{\theta\in\Theta}\int_{E}\ln p\left(y;\theta\right)P_m\left(\mathrm{d}y|B_m\left(x\right)\right)\quad\textit{and}\quad\theta_0\left(x\right):=\arg\inf_{\theta\in\Theta}\mathbb{E}\left[\ln p\left(X_n;\theta\right)|X_{n-1}=x\right].$$

3.2. Examples of general state space

We give a simple example of a general state space, in particular, we shall discuss the case $E\subseteq\mathbb{R}^{\mathbb{N}}$ equipped with the metric $d_\infty\left(x,y\right)=\sum_{i=1}^{\infty}2^{-i}f\left(d\left(x_i,y_i\right)\right)$ where $x_i,y_i\in\mathbb{R}$, $f\left(t\right)=t/\left(1+t\right)$ and d is any metric topologically equivalent to the Euclidean norm. Then, $\mathbb{R}^{\mathbb{N}}$ is metrizable by d_∞ ([17], Proposition 2.4.4) which turns $\mathbb{R}^{\mathbb{N}}$ into a separable metric space, so that the sigma algebra of its subsets is countably generated and, mutatis mutandis, the results of the paper can be derived in this more general framework, where the conditioning sets are balls of d_∞-radius r_m. In this case, Theorem 6 does not hold uniformly because the collection of sets $\left\{y\in E\subseteq\mathbb{R}^{\mathbb{N}}:y\le s\right\}$, $s\in E$, does not have a finite bracketing number (see Definition 30 for the exact meaning in this context). Nevertheless, let $\lambda:E\to\mathbb{R}^K$ ($K\ge1$ but finite). We can then estimate uniformly

$$\Pr\left(\lambda\left(X_n\right)\le s|X_{n-1}=x\right)$$

where $s\in\lambda\left(E\right)$ because the collection $\left\{y\in\lambda\left(E\right)\subseteq\mathbb{R}^K:y\le s\right\}$, $s\in\lambda\left(E\right)$, has a finite bracketing number. Hence, if we are only interested in the restriction of X in $\lambda\left(E\right)$, we can allow for larger classes of functions as discussed next. In this case, the functions in $\mathfrak{F}$ are functions from $\lambda\left(E\right)$ and not from E.

Another example that has attracted recent attention is the case of functional data (e.g. [2]). The results given here apply to this setting as well. Suppose $\left(X_i\right)_{i\in\mathbb{N}}=\left(X_i\left(u\right);u\in\mathcal{U}\right)_{i\in\mathbb{N}}$ where $\mathcal{U}$ is a compact set and E is a set of uniformly equicontinuous functions with values in $\mathbb{R}$. By the Arzela Ascoli Theorem (e.g. [17], Theorem 2.4.7) E is a totally bounded metric space under the uniform norm, say $d_{\sup}$. Hence, E has a countable base and its Borel sigma algebra is countably generated.

3.3. Extensions of Condition 9

Condition 9 ii. restricts attention to π_x-a.s. equicontinuous families of functions. However, for $E\subseteq\mathbb{R}^K$, Theorem 6 holds uniformly for $I\left\{y\in E:y\le s\right\}$, $s\in E$, which is not continuous. Hence, as mentioned in Remark 11, it is clear that we could consider larger classes of functions (though in the statement of the results we refrained to do so to avoid extra notation). We recall the following definition.

Definition 24. A function f on E is of Hardy bounded variation (BV) (e.g. [18,19]) if for any $y \in E \subseteq \mathbb{R}^K$,

$$f(y) = \mu_1(\{s \in E : s \leq y\}) - \mu_2(\{s \in E : s \leq y\})$$

where μ_1 and μ_2 are Radon measures. (Note that the Radon measure of a compact set is finite.)

Remark 25. In one dimension this is the usual definition of bounded variation. In higher dimensions, there is no unique way to define bounded variation (e.g. [18]), though the usual definition is different (e.g. [20]).

Then, we note the following.

Corollary 26. *Suppose BV_b is the class of uniformly bounded functions in BV. Under the Conditions of Theorem 6,*

$$\sup_{f \in BV_b} \left| \hat{P}f(B_m(x)) - Pf(x) \right| \xrightarrow{a.s.} 0,$$

for any $x \in E$.

Proof. Let $\mathfrak{M}_b$ be the class of bounded monotone increasing functions in each argument with domain E. It is sufficient to prove uniform convergence in $\mathfrak{M}_b$. Hence, by Lemma 10 in [21] deduce that

$$\sup_{f \in \mathfrak{M}_b} \left| \hat{P}f(B_m(x)) - Pf(x) \right| \xrightarrow{a.s.} 0 \quad \text{if and only if} \quad \sup_{s \in E} \left| \hat{P}(s|B_m(x)) - P(s|x) \right| \xrightarrow{a.s.} 0$$

and the result is proved. ■

For definiteness let $\mathfrak{E}_b$ be an arbitrary, but fixed, family of uniformly bounded equicontinuous functions. Note that by equicontinuity, each element in $\mathfrak{E}_b$ can be turned into a Lipschitz function under the metric

$$d(x, y) := \sup_{f \in \mathfrak{E}_b} |f(x) - f(y)|$$

for each $x, y \in E$ (see the proof of Corollary 11.3.4 in [17]). This shows that $\mathfrak{E}_b$ may contain many functions of interest on top of Lipschitz functions under more standard metrics. However, by Corollary 26 we may further increase the set of functions allowed by Condition 9 ii. to $\mathfrak{F} \subseteq \mathfrak{E}_b \cup BV_b$. Note that while the intersection of $\mathfrak{E}_b$ and BV_b is not empty, it is not possible to establish an inclusion of one family into another. In fact there are uniformly continuous functions that are not of bounded variation (e.g. $f(y) = y \sin(1/y)$ for $y \in (0, 2\pi]$, 0 elsewhere, is not in BV_b). Clearly, $f(y) = \{s \in E : s \leq y\}$ is in BV_b but not in $\mathfrak{E}_b$. Hence $\mathfrak{E}_b \cup BV_b$ is fairly rich and we may allow for convex combinations of these functions as well (e.g. [22] Ch. 2.10 for details). A tail condition as in Condition 9 i. allows us to truncate so that we can avoid the uniform boundedness condition (see Lemma 34 in Section 4).

3.4. Generalizations of Theorem 20

Theorem 20 is based on a simple martingale approximation. The proof reduces to showing a CLT for

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left(1 - \mathbb{E}_{T(i)}\right) f\left(X\left(T_{B_m}(i) + 1\right)\right).$$

Hence, the result can be extended in two directions. We can allow for finite collections of functions $\mathfrak{F}$ that are not necessarily uniformly bounded, using the Lindeberg type of conditions. These are well known (e.g. [23]). On the other hand, we can allow $\mathfrak{F}$ to be uncountable at the cost of imposing smoothness conditions on the functions in $\mathfrak{F}$. This would rule out the conditional empirical distribution function. Here we provide some results when $\mathfrak{F}$ is uncountable and its elements are not necessarily smooth functions. One approach is to use an argument based on a uniform central limit theorem for families of uniformly bounded martingales (e.g. [24]). Another approach is to use results based on HRMC and small sets. We give the details in the latter case when $E = \mathbb{R}$.

Condition 27. *There is a set $C \subset \mathbb{R}$ and a probability measure ν with support C such that for some $s \in (0, 1)$, and any positive integer n*

$$P^n(x, A) \geq \left(1 - s^n\right) \nu(A)$$

for any $x \in C$ and $A \subseteq C$.

Condition 27 is the standard minorization condition in MC theory (see [11] for details, and [29] for an application in a context similar to the one of this paper). We shall restrict attention to BV functions with domain in $\mathbb{R}$ (see Definition 24). Using the representation of a BV function f as given in Definition 24, define the total variation over $\mathbb{R}$: $\|f\|_{TV} := \mu_1(\mathbb{R}) + \mu_2(\mathbb{R})$. If $\|f\|_{TV} < \infty$, then $f \in BV$ and has finite support (a BV function only needs to have finite total variation over bounded subsets of $\mathbb{R}$). Define BV_1 to be the class of BV functions such that $\|f\|_{TV} \leq 1$. Let $\mathfrak{F} \subseteq BV_1$ equipped with the seminorm $\|\bullet\|_{TV}$. The metric entropy $\mathcal{H}(\epsilon, \mathfrak{F}, \|\bullet\|_{TV})$ is the logarithm of the minimum number of balls of $\|\bullet\|_{TV}$-radius ϵ needed to cover $\mathfrak{F}$ (e.g. [22], Ch. 2.1, for further details).

Condition 28. *$\mathfrak{F} \subseteq BV_1$ and*

$$\int_0^1 \mathcal{H}\left(\epsilon, \mathfrak{F}, \|\bullet\|_{TV}\right) \mathrm{d}\epsilon < \infty.$$

The above display is the Dudley metric entropy integral and the application with the total variation seminorm is the taken from [25]. We have the following uniform central limit theorem.

Theorem 29. *Under Conditions 1, 17, 18, 27 and 28,*

$$\sqrt{m}\left(\hat{P}_m f\left(B_m(x)\right) - Pf(x)\right) \to \mathbb{G}_x(f)$$

weakly, where $(\mathbb{G}_x(f))_{f\in\mathfrak{F}}$ is a mean zero Gaussian process with covariance function

$$\mathbb{E}\mathbb{G}_x(f)\,\mathbb{G}_x(g) = P(fg)(x) - Pf(x)\,Pg(x), \quad f, g \in \mathfrak{F}.$$

4. Proofs

Some of the material included in the proofs is known, but is still included for the sake of completeness and ease of readability. We recall the definition of bracketing numbers (e.g. [22] for more details) to be used in the present context.

Definition 30. For measurable functions l and u, the bracket $[l, u]$ is the set of all functions f such that $l \le f \le u$ and an $L_p(\pi_x)$ ϵ-bracket is a bracket such that $\left[P|u-l|^p(x)\right]^{1/p} \le \epsilon$. The minimal number of $L_p(\pi_x)$ ϵ-brackets needed to cover a set $\mathfrak{F}$ is called the bracketing number.

We can now turn to the proof of the results.

4.1. Proof of Theorem 6

By the triangle inequality,

$$\left|\hat{P}_m\left(s|B(x, r_m)\right) - P(s|x)\right| \le \left|\hat{P}_m(s|B_m) - P(s|B_m)\right| + \left|P(s|B_m) - P(s|x)\right|$$

where the first term on the r.h.s. is the estimation error, while the second term is the approximation error. When $E = \mathbb{R}^K$, we take $\sup_{s\in E}$ on both sides of the above equality. Then, convergence of the estimation error follows from the following.

Lemma 31. *Under Condition 1, for any $B \subset E$, such that $\pi(B) < \infty$,*

$$\hat{P}_m(s|B) \xrightarrow{a.s.} P(s|B),$$

for any $s \in E$ and if $E \subseteq \mathbb{R}^K$,

$$\sup_{s\in E}\left|\hat{P}_m(s|B) - P(s|B)\right| \xrightarrow{a.s.} 0,$$

as $m \to \infty$.

Proof. Note that

$$\begin{aligned}\hat{P}_m(s|B) &= \frac{1}{m}\sum_{i=1}^{m} I\left\{X\left(T_B(i)+1\right) \le s\right\} \\ &= \frac{\sum_{i=1}^{n} I\{X_i \in B\}}{m}\, \frac{\sum_{i=1}^{n} I\{X_i \in B, X_{i+1} \le s\}}{\sum_{i=1}^{n} I\{X_i \in B\}}\end{aligned}$$

where n is such that

$$m = \sum_{i=1}^{n} I\{X_i \in B\}.$$

Clearly, given m, n is random, and given n, m is random, but in any case one goes to infinity a.s. if the other does. Hence,

$$\frac{\sum_{i=1}^{n} I\{X_i \in B, X_{i+1} \le s\}}{\sum_{i=1}^{n} I\{X_i \in B\}} \xrightarrow{a.s.} P(s|B)$$

by Proposition 8.2.7(3.) in [26]. This implies the pointwise convergence result. To obtain uniform convergence when $E \subseteq \mathbb{R}^K$, note that we can find a finite number S of bracketing functions $(I\{X_n \le y_s\}, s = 1, \ldots, S)$ for the indicator function of sets of the form $\{y \in E \subseteq \mathbb{R}^K, y \le s\}$ (K finite) such that

$$\mathbb{E}(|I\{X_n \le y_{s+1}\} - I\{X_n \le y_s\}| \,|X_{n-1} = x) \le \epsilon,$$

where $y_{s+1} > y_s$. Hence, the convergence is also uniform (e.g. Theorem 2.4.1 in [22], for further details). ■

We now consider the approximation error.

Lemma 32. *Set $B_m := B(x, r_m)$. By Conditions 2 and 4*

$$|P(s|B_m) - P(s|x)| \to 0$$

and if $E \subseteq \mathbb{R}^K$, the convergence holds uniformly in $s \in E$.

$$\sup_{s\in E} |P(s|B_m) - P(s|x)| \to 0.$$

Proof. Recall that $Pf(B_m) := \int_{B_m}\int_E f(y) P(x, dy)[\pi(dx)/\pi(B_m)]$. Then,

$$\begin{aligned} |P(s|B_m) - P(s|x)| &= \left|\int_{B_m} [P(s|y) - P(s|x)] \frac{\pi(dy)}{\pi(B_m)}\right| \\ &\le \sup_{y\in B_m} |P(s|y) - P(s|x)| \\ &\to 0 \end{aligned}$$

by Condition 2 as $B_m \to \{x\}$. If Condition 2 does not hold, but $y \in \mathbb{R}^K$, as mentioned in Remark, by differentiation of integrals,

$$\frac{1}{\pi(B_m)}\int_{B_m} P(s|y)\pi(dy) \to P(s|x)$$

for π-almost all x because π is a Radon measure (e.g. locally finite). When $E \subseteq \mathbb{R}^K$, using a finite number of bracketing functions for the indicator function of sets $\{y \in E : y \le s\}, s \in E$, as in Lemma 31,

$$\sup_{s\in E} |P(s|B_m) - P(s|x)| = \sup_{s\in E} |\Pr(X_n \le s|X_{n-1} \in B(x, r_m)) - \Pr(X_n \le s|X_{n-1} = x)| \xrightarrow{a.s.} 0. \quad ■$$

The theorem is proved.

4.2. Proof of corollaries

Proof (*Proof of* Corollary 12). We use a standard truncation argument (e.g. proof of Theorem 1.11.3 in [22]). Set $f^b := fI_{\{|f|>b\}}$ and $f_b := fI_{\{|f|\le b\}}$. Then,

$$\begin{aligned} \sup_{f\in\mathfrak{F}} \left|\hat{P}_m f(B_m(x)) - Pf(x)\right| &\le \sup_{f\in\mathfrak{F}} \left|\hat{P}_m f_b(B_m(x)) - Pf_b(x)\right| + \sup_{f\in\mathfrak{F}} \left|\hat{P}_m \left|f^b\right|(B_m(x)) + P\left|f^b\right|(x)\right| \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

Since $f_b \in \mathfrak{C}_b$ by ii. in Condition 9, $\mathrm{I} \xrightarrow{a.s.} 0$ using Lemma 33 stated next.

Lemma 33. *Let $\mathfrak{C}_b$ be a family of π_x-a.s. uniformly bounded and equicontinuous functions. Under the conditions of Theorem 6,*

$$\sup_{f\in\mathfrak{C}_b} \left|\hat{P}_m f(B_m(x)) - Pf(x)\right| \xrightarrow{a.s.} 0.$$

Proof. By Theorem 6, $\hat{P}_m(s|B_m(x))$ converges a.s. to $P(s|x)$. Then, uniform convergence in $\mathfrak{C}_b$ follows by Corollary 11.3.4 in [17]. ■

Since the envelope of $\mathfrak{F}$ satisfies suitable moment conditions, II $\overset{a.s.}{\leq} \epsilon$ where ϵ is arbitrary for b large enough using the Lemma 34 stated next.

Lemma 34. *Suppose $\mathfrak{F}$ satisfies* i. *in Condition 9. Then, for any $\epsilon > 0$, there is a large enough b such that*

$$\sup_{f \in \mathfrak{F}} \left| \hat{P}_m \left| f I_{\{|f|>b\}} \right| (B_m(x)) + P \left| f I_{\{|f|>b\}} \right| (x) \right| \overset{a.s.}{\leq} \epsilon.$$

Proof. Set $F^b := F I_{\{F>b\}}$, where F is the envelope of $\mathfrak{F}$. By the triangle inequality,

$$\begin{aligned} \left| \hat{P}_m F^b (B_m(x)) + P F^b (x) \right| &\leq \left| \hat{P}_m F^b (B_m(x)) - P F^b (B(x)) \right| + \left| P F^b (B(x)) + P F^b (x) \right| \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

Condition 9 i. allows us to apply the convergence result in [26] cited in the proof of Lemma 31; hence I$\overset{a.s.}{\to} 0$. By Condition 9 i, since $B_m(x) \to \{x\}$, $PF^p(x) \leq \limsup_m PF^p(B_m(x)) < \infty$ implies II $\leq \epsilon$, for any $\epsilon > 0$, by suitable choice of b. Noting that

$$\sup_{f \in \mathfrak{F}} \hat{P}_m \left| f I_{\{|f|>b\}} \right| (B_m(x)) \leq \hat{P}_m F^b (B_m(x)),$$

and similarly for P, the result follows. ■

Hence, the corollary is proved. ■

Proof (*Proof of* Corollary 15). The proof can be deduced from the proof of Lemma 35. ■

4.3. *Proof of* Theorem 16

By Harris recurrence, a.s., $m \to \infty$ if and only if $n \to \infty$. Hence, by Lemma 35 (next), $\rho\left(\hat{f}_{m,n}, f_n\right) \overset{p}{\to} 0$ conditioning on $X_0 = x$ as $n \to \infty$.

Lemma 35. *Suppose $(\rho, \mathfrak{F})$ is a metric space. Under Conditions 1, 2, 4, 9 and 14, conditioning on $X_0 = x$,*

$$\rho\left(\hat{f}_{m,n}, f_n\right) \overset{p}{\to} 0.$$

Proof. Note that $f_n := f_n(X_{n-1})$ and $\hat{f}_{m,n} := \hat{f}_m(X_{n-1})$ are random, as they depend on X_{n-1} (dependence of $\hat{f}_m$ on the sample values is suppressed for ease of notation). Let $G^{(n)} = G^{(n)}(X_{n-1})$ be an arbitrary open set that contains f_n and let $\left[G^{(n)}\right]^c$ be its complement. It is enough to show that

$$\mathrm{I} := \Pr\left(f_n \in G^{(n)}, \hat{f}_n \in \left[G^{(n)}\right]^c\right) = o(1),$$

as $G^{(n)}$ is arbitrary. To this end note that

$$\mathrm{I} = \Pr\left(\inf_{f \in \left[G^{(n)}\right]^c} \hat{P}_m f(B_m(X_{n-1})) \leq \inf_{f \in G^{(n)}} \hat{P}_m f(B_m(X_{n-1})), f_n \in G^{(n)}\right)$$

because the infimum of $\hat{P}_m f(B_m(X_{n-1}))$ is attained in $\left[G^{(n)}\right]^c$. Moreover, note that for any set $A \subseteq \mathfrak{F}$

$$\begin{aligned} &\inf_{f \in A} Pf(X_{n-1}) - \sup_{f \in A} \left| \hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1}) \right| \\ &\quad \leq \inf_{f \in A} \hat{P}_m f(B_m(X_{n-1})) \leq \inf_{f \in A} Pf(X_{n-1}) + \sup_{f \in A} \left| \hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1}) \right|. \end{aligned}$$

Define

$$R_n := \sup_{f \in G^{(n)}} \left| \hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1}) \right|,$$

and

$$R_n' := \sup_{f \in \left[G^{(n)}\right]^c} \left| \hat{P}_m f(B_m(X_{n-1})) - Pf(X_{n-1}) \right|.$$

Then,

$$\begin{aligned}
\mathrm{I} &\leq \Pr\left(\inf_{f\in\left[G^{(n)}\right]^c} Pf\left(X_{n-1}\right) \leq \inf_{f\in G^{(n)}} Pf\left(X_{n-1}\right) + R_n + R_n', f_n \in G^{(n)}\right)\\
&\leq \Pr\left(\inf_{f\in\left[G^{(n)}\right]^c} Pf\left(X_{n-1}\right) \leq \inf_{f\in G^{(n)}} Pf\left(X_{n-1}\right) + 2\epsilon, f_n \in G^{(n)}\right)\\
&\quad + \int_E \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, \mathrm{d}x_{n-1}\right) + \int_E \Pr\left(R_n' \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, \mathrm{d}x_{n-1}\right)\\
&= \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

Since ϵ is arbitrary, by Condition 14, $\mathrm{II} = 0$ because either $f_n \in \left[G^{(n)}\right]^c$ or $f_n \in G^{(n)}$. By Corollary 12, $\Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) \to 0$ for any $x_{n-1} \in E$. Moreover,

$$\int_E \Pr\left(R_n \geq \epsilon | X_{n-1} = x_{n-1}\right) P^{n-1}\left(x, \mathrm{d}x_{n-1}\right) \leq \int_E 1 P^{n-1}\left(x, \mathrm{d}x_{n-1}\right) \leq P^{n-1}\left(x, E\right) = 1.$$

Hence $\mathrm{III} \to 0$ by the Dominated Convergence Theorem. An identical argument shows that $\mathrm{IV} \to 0$ as well. ■

Then, the Cesaro sum of $\rho\left(\hat{f}_{m,n}, f_n\right)$ goes to zero by an application of Lemma 36.

Lemma 36. *Suppose $\left(Z_n\right)_{n\in\mathbb{N}}$ is a sequence of uniformly integrable positive random elements such that $Z_n \xrightarrow{p} 0$. Then,*

$$\frac{1}{N}\sum_{n=1}^{N} Z_n \to 0 \quad in\ L_1.$$

Proof. The proof is included for ease of reference. For any $N' < N$,

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}Z_n &= \frac{1}{N}\sum_{n=1}^{N'} \mathbb{E}Z_n + \frac{1}{N}\sum_{n=N'}^{N} \mathbb{E}Z_n\\
&\leq \max_{1\leq n\leq N'} \frac{N'}{N}\mathbb{E}Z_n + \max_{N'\leq n\leq N} \mathbb{E}Z_n\\
&= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Let $N' = o\left(N\right)$, so that by uniform integrability $\mathrm{I} \to 0$. Recall that convergence in probability plus uniform integrability is equivalent to convergence in L_1 (e.g. [27], Theorem 21.2), so that $\mathbb{E}Z_n \to 0$. Letting $N' \to \infty$ as $N \to \infty$,

$$\lim_{N'} \mathrm{II} \leq \lim_{N'} \sup_{n\geq N'} \mathbb{E}Z_n \to 0$$

because $Z_n \xrightarrow{p} 0$. ■

Hence, the theorem is proved.

4.4. *Proof of* Theorems 20 *and* 29

Proof (*Proof of* Theorem 20). At first, note that

$$\mathbb{E}_{T(i)} f\left(X\left(T_{B_m}\left(i\right) + 1\right)\right) = Pf\left(x_i\right)$$

for some $x_i \in B_m\left(x\right)$. Hence,

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\mathbb{E}_{T(i)} f\left(X\left(T_{B_m}\left(i\right) + 1\right)\right) - Pf\left(x\right)\right] &\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left|Pf\left(x_i\right) - Pf\left(x\right)\right|\\
&\lesssim \sqrt{m} r_m^{\alpha} \to 0
\end{aligned}$$

by Condition 17. By the above display, it is enough to show convergence in distribution of

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left(1 - \mathbb{E}_{T(i)}\right) f\left(X\left(T_{B_m}\left(i\right) + 1\right)\right), \tag{5}$$

where

$$\left\{\left(1-\mathbb{E}_{T(i)}\right) f\left(X\left(T_{B_m}(i)+1\right)\right): f \in \mathfrak{F}\right\}$$

is a finite family of uniformly bounded martingale differences. Then, the result follows by an application of the central limit theorem for martingales (e.g. Theorem 2.3 in [23]) and the Cramer Wold device. ■

Proof of Theorem 29 (*Proof of* Theorem 29). From the proof of Theorem 20, it is enough to show weak convergence of (5). Condition 28 (i.e. the metric integral) implies $\mathscr{H}\left(\epsilon, \mathfrak{F}, \|\bullet\|_{TV}\right) < \infty$ for any $\epsilon > 0$ ($\mathscr{H}\left(\epsilon, \mathfrak{F}, \|\bullet\|_{TV}\right)$ is decreasing in ϵ), so that, by definition, $\mathfrak{F}$ is totally bounded. Hence, to show weak convergence we only need to show finite dimensional (fidi) convergence plus stochastic equicontinuity of (5) (e.g. [22] Theorem 1.5.7). Fidi convergence follows from Theorem 20 and we only need to show stochastic equicontinuity. Lemma A1 in [29] shows that under Condition 27, $\left(X\left(T_{B_m}(i)\right)\right)_{i>0}$ has an invariant distribution and is phi mixing with geometrically decaying mixing coefficients

$$\varphi_i := \sup_{A \subseteq \mathbb{R}, x \in \mathbb{R}} \left|\Pr\left(X\left(T_{B_m}(i+j)\right) \in A \middle| X\left(T_{B_m}(j)\right)=x\right)-\Pr\left(X\left(T_{B_m}(i+j)\right) \in A\right)\right|,$$

for any $B_m \subset C$, hence for any $m > 0$ (C as in Condition 27). We have suppressed dependence on m in the mixing coefficient φ_i and used the fact that $X\left(T_{B_m}(i+j)\right)$ depends on $X\left(T_{B_m}(j)\right), X\left(T_{B_m}(j-1)\right), X\left(T_{B_m}(j-2)\right), \ldots$, only through $X\left(T_{B_m}(j)\right)$, by the strong Markov property. By Proposition 7.6 in [28], we have the following representation for an MC:

$$X_{i+1} = h\left(X_i, \zeta_{i+1}\right),$$

for some measurable function h and an iid sequence of uniform $(0, 1)$ random variables $(\zeta_i)_{i>0}$. Since for any fixed $z \in (0, 1)$, $h\left(X\left(T_{B_m}(i)\right), z\right)$ is a measurable transformation of $X\left(T_{B_m}(i)\right)$, from the definition of φ_i it can be deduced that also $\left(X\left(T_{B_m}(i)+1\right)\right)_{i>0}$ is phi mixing with geometrically decaying mixing coefficients. Hence stochastic equicontinuity follows from Corollary 4 in [25]. ■

Acknowledgments

The author would like to thank Paul Doukhan and Emmanuel Guerre for some very useful information that he used in the proof of Theorem 29, Brendan Beare for a discussion about functions of bounded variation, an Editor and a Referee for comments that led to improvements both in content and presentation.

References

[1] P.M. Robinson, Nonparametric estimators for time series, Journal of Time Series Analysis 4 (1983) 185–207.
[2] E. Masry, Nonparametric regression for dependent functional data: Asymptotic normality, Stochastic Processes and their Applications 115 (1) (2005) 155–177.
[3] P. Ango Nze, P. Bühlmann, P. Doukhan, Weak dependence beyond mixing and asymptotics for nonparametric regression, Annals of Statistics 30 (2002) 397–430.
[4] P. Doukhan, Mixing: Properties and Examples, in: Lecture Notes in Statistics, vol. 85, Springer, New York, 1994.
[5] P. Doukhan, S. Louhichi, A new weak dependence condition and applications to moment inequalities, Stochastic Processes and Applications 84 (1999) 313–342.
[6] P. Ango Nze, P. Doukhan, Weak dependence: Models and applications to econometrics, Econometric Theory 20 (2004) 995–1045.
[7] Sid Yakowitz, Nearest neighbor regression estimation for null-recurrent markov time series, Stochastic Processes and their Applications 48 (1993) 311–318.
[8] H.A. Karlsen, D. Tjøstheim, Nonparametric estimation in null recurrent time series, Annals of Statistics 29 (2001) 372–416.
[9] J.L. Horowitz, Bootstrap methods for Markov processes, Econometrica 71 (2003) 1049–1082.
[10] P. Bertail, S. Clémençon, Regenerative block bootstrap for Markov chains, Bernoulli 12 (2006) 689–712.
[11] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer, London, 1993.
[12] M. Babillot, P. Bougerol, L. Elie, The random difference equation $X_n = A_n X_{n-1} + B_n$ in the critical case, Annals of Probability 25 (1997) 478–493.
[13] L. de Haan, S. Resnick, H. Rootz én, C. de Vries, Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes, Stochastic Processes and their Applications 32 (1989) 213–224.
[14] N.H. Bingham, C.M. Goldie, J.L. Teugels, Regular Variation, Cambridge University Press, Cambridge, 1987.
[15] X. Chen, How often does a harris recurrent markov chain recur? Annals of Probability 27 (1999) 1324–1346.
[16] K.-J. Van Garderen, Curved exponential models in statistics, Econometric Theory 13 (1997) 771–790.
[17] R.M. Dudley, Real Analysis and Probability, Cambridge University Press, Cambridge, 2002.
[18] J.A. Clarkson, C.R. Adams, On definitions of bounded variation for functions of two variables, Transactions of the American Mathematical Society 35 (1933) 824–854.
[19] B. Lenze, On the points of regularity of multivariate functions of bounded variation, Real Analysis Exchange 29 (2003) 646–656.
[20] W. Ziemer, Weakly Differentiable Functions, Springer, New York, 1989.
[21] Sancetta, Weak convergence of Laws on $\mathbb{R}^K$ with common marginals, Journal of Theoretical Probability 20 (2007) 371–380. Downloadable: http://arxiv.org/abs/math.PR/0606462.
[22] A. Van der Vaart, J.A. Wellner, Weak convergence of empirical processes, in: Springer Series in Statistics, Springer, New York, 2000.
[23] D.L. McLeish, Dependent central limit theorems and invariance principles, Annals of Probability 2 (1974) 620–628.
[24] S. Levental, A uniform CLT for uniformly bounded families of martingale differences, Journal of Theoretical Probability 3 (1989) 271–287.
[25] J. Dedecker, C. Prieur, New dependence coefficients. Examples and applications to statistics, Probability Theory and Related Fields 132 (2005) 203–236.
[26] M. Duflo, Random Iterative Models, Springer, Berlin, 1997.
[27] L.C.G. Rogers, D. Williams, Diffusions, Markov Processes and Martingales, Cambridge University Press, Cambridge, 2000.
[28] O. Kallenberg, Foundations of Modern Probability, Springer, Berlin, 1997.
[29] E. Guerre, Design-adaptive pointwise nonparametric regression estimation for recurrent markov time series. INSEE D.P. n. 2004–22, 2004.
[30] Luc Devroye, On the almost everywhere convergence of nonparametric regression function estimates, Annals of Statistics 9 (1981) 1310–1319.